# Reverse Engineering Psychologically Valid Facial Expressions of Emotion into Social Robots

Chaona Chen*, Oliver G.B. Garrod*, Jiayu Zhan*, Jonas Beskow† Philippe G. Schyns* and Rachael E. Jack*
*Institute of Neuroscience and Psychology, University of Glasgow, G12 8QB, Scotland, UK
†Furhat Robotics, 11428 Stockholm, Sweden
Email: rachael.jack@glasgow.ac.uk

***Abstract*—Social robots are now part of human society, destined for schools, hospitals, and homes to perform a variety of tasks. To engage their human users, social robots must be equipped with the essential social skill of facial expression communication. Yet, even state-of-the-art social robots are limited in this ability because they often rely on a restricted set of facial expressions derived from theory with well-known limitations such as lacking naturalistic dynamics. With no agreed methodology to objectively engineer a broader variance of more psychologically impactful facial expressions into the social robots' repertoire, human-robot interactions remain restricted. Here, we address this generic challenge with new methodologies that can reverse-engineer dynamic facial expressions into a social robot head. Our data-driven, user-centered approach, which combines human perception with psychophysical methods, produced highly recognizable and human-like dynamic facial expressions of the six classic emotions that generally outperformed state-of-art social robot facial expressions. Our data demonstrates the feasibility of our method applied to social robotics and highlights the benefits of using a data-driven approach that puts human users as central to deriving facial expressions for social robots. We also discuss future work to reverse-engineer a wider range of socially relevant facial expressions including conversational messages (e.g., interest, confusion) and personality traits (e.g., trustworthiness, attractiveness). Together, our results highlight the key role that psychology must continue to play in the design of social robots.**

***Keywords*-facial expressions; data-driven methods; social robots; human social perception; psychophysical methods**

## I. Introduction

Teaching medical staff good bedside manners involves realistic displays of fear, anger, and sadness, whereas household companion robots must look friendly and trustworthy. To equip social robots with these complex skills of social communication, social roboticists often turn to psychologists to understand which facial expressions elicit these judgments in humans, particularly of emotions such as anger, sadness, and happiness. One popular approach is to use a set of prototypical facial expressions derived from theory (e.g., [1]) and install them into social robots (e.g., [2-4]). However, this approach limits the expressive capacity of social robots because these facial expressions are derived from constrained theory-driven methods that focus primarily on a small and specific set of static facial expressions selected by researchers in a top-down manner (e.g., [5]). In addition to lacking important dynamical information such as the temporal order of face movements (e.g., [6]), and being recognized mostly by Westerners (e.g., [7-9]), such an approach cannot adequately represent the variance of facial expressions that communicate emotions to human observers. Consequently, social robots remain limited in their ability to generate the wide range of complex facial expressions required to engage their human users in the nuances of everyday social interaction (e.g., [10]).

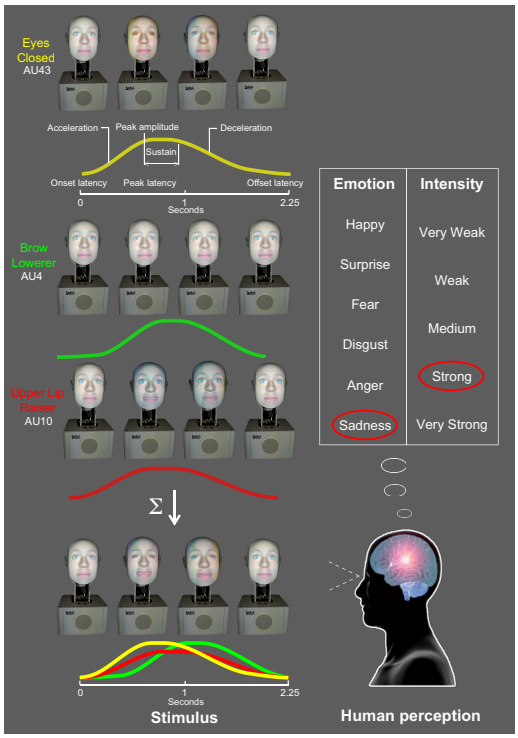


Figure 1. An illustration of stimulus generation and task procedure.

This work was supported by The Economic and Social Research Council and Medical Research Council (United Kingdom; ESRC/MRC-060-25-0010), British Academy (BA SG171783), Wellcome Trust (107802/Z/15/Z) and Multidisciplinary University Research Initiative (MURI)//Engineering and Physical Sciences Research Council (EP/N019261/1).

## II. RELATED WORK

### A. Data-driven Modelling of Dynamic Facial Expressions

To alleviate these constraints and address subsequent knowledge gaps, new data-driven methods can now mathematically model dynamic facial expressions (e.g., [11]) using a dynamic face movement generator [12], reverse correlation [13], and subjective human perception (see [14] for an overview). Rather than selecting facial expressions top-down, this approach agnostically samples a broad range of facial expressions, tests them against human perception, and then reverse engineers the specific dynamic facial expression patterns that elicit the perception of different emotions in human users (e.g., [12, 15]). The resulting facial expression models uniquely detail how individual face movements called Action Units (AUs) [1], such as Eyebrow Raiser (AU1-2), Nose Wrinkler (AU9), and Jaw Drop (AU26), are activated over time using several temporal parameters (e.g., acceleration, peak amplitude, sustainment) and thus provide a precise information ontology for facial expression communication. This approach can therefore objectively establish the correspondence between facial expression patterns and human emotion perception, delivering results that are psychologically and culturally meaningful, and placing the human user as central to deriving facial expression signals (see [16] for a review). Here, we use one such data-driven method to reverse engineer, for the first time, a set of dynamic facial expressions of emotion directly on a popular social robot head (Furhat, see www.furhatrobotics.com). We also evaluated the recognition accuracy and humanlike-ness of these facial expressions and compared their performance to those currently installed in the social robot head.

## III. METHOD

### A. Modelling Dynamic Facial Expressions of Emotion

Fig. 1 illustrates the stimulus generation and task procedure using one illustrative trial. On each experimental trial, we randomly selected a subset of AUs from a core set of 41 AUs using a binomial distribution (parameters set to n = 3, P = 0.6). For example, in Fig. 1 the selection comprises Eyes Closed (AU43) color-coded in yellow, Brow Lowerer (AU4) in green, and Upper Lip Raiser (AU10) in red. For each randomly chosen AUs separately, we assigned a random movement by selecting random values for each of seven temporal parameters (onset latency, acceleration, peak amplitude, sustainment, peak latency, deceleration, offset latency; see labels illustrating the yellow curve in Fig. 1) using a uniform distribution over the interval [0, 1]. To activate each AU, we used a cubic Hermite spline (a smooth interpolation curve with 6 points controlling the shape of the curve) and constrained the dynamics using a set of temporal equations that specify how each of the temporal parameters is mathematically related to each of the 6 control points (see [12] for full details). We then sampled the resulting smooth curve over 60 time frames to produce the facial animation. We combined these randomly activated AUs to create a random facial animation of 2.25-seconds duration (bottom row of Fig. 1 shows an example using four snapshots across time). We displayed each animation on one of the 7 face textures available on the social robot head ('Default,''Male,' 'Female,' 'Obama,' 'iRobot,' 'Gabriel,' and 'Avatar') and back-projected the stimulus on to the robot's plastic face mask. The participant viewed the random facial animation, and if it formed a pattern that correlated with their prior knowledge of one of the six emotions ('happy,' 'surprise,' 'fear,' 'disgust,' 'anger' or 'sadness'), they categorized it accordingly (e.g., in Fig. 1, 'sadness') and rated its intensity on a 5-point scale (e.g., in Fig. 1, 'strong'). Otherwise, if none of the emotion labels accurately described the facial expression, the participant selected 'other.' We played each animation only once. Participants responded using a Graphic User Interface (GUI) displayed on a 19-inch flat panel Dell monitor positioned next to the social robot head, and had unlimited time to respond. We generated a total of 3605 such facial animations and pseudo-randomly assigned each to one of the 7 face textures (515 animations per texture) for each participant separately. We blocked trials by face texture and randomized the order of the blocks for each participant. All face stimuli (size 22.5 cm $\times$ 16 cm) appeared in the participant's central visual field at a constant viewing distance of 90 cm using a chin rest. Stimuli subtended 14.25° (vertical) and 10.16° (horizontal) of visual angle, which reflects the average size of a human face [17] during natural social interaction [18]. We used Matlab 2016a to display the GUI and record responses.

We recruited 2 white Westerners (1 female, mean age 24 years, SD = 2.83 years) with minimal exposure to and engagement with other cultures [19] as assessed by questionnaire (see *Supporting Information, Observer Questionnaire* in [11] for full details). All participants had normal or corrected-to-normal vision, were free from any emotion related atypicalities (e.g. Autism Spectrum Disorder, depression, anxiety), learning difficulties (e.g. dyslexia), synesthesia, and disorders of face perception (e.g. prosopagnosia) as per self-report. Each participant gave written informed consent, and received a standard rate of £6 per hour for their participation. The Ethics Committee of the College of Science and Engineering, University of Glasgow provided ethical approval (Ref No: 300160186).

### B. Building a Facial Movement Vocabulary for the Social Robot

To equip the social robot head with the ability to generate these AU movements, we transferred our existing method of arbitrary AU combination synthesis, which is based on a library of 41 human performance-captured AUs each with 7 temporal parameters, by transferring the AU shape deviation data between the two different mesh topologies. Thus, we

augmented the social robot's facial movement vocabulary from a set of 7 preset prototypical facial expressions (2 'happy,' 1 'surprised,' 1 'fear,' 1 'disgust,' 1 'anger' and 1 'sadness'), plus 6 eyebrow modifiers, and 2 blink modifiers [20], to include an extensive library of dynamic AUs and their combinations. This technical contribution provides the new advance of using powerful data-driven reverse-engineering methods on this social robot head.

### C. Facial Expression Model Fitting

To identify the dynamic AUs that are significantly correlated with the human perception of each emotion, we used an established model fitting procedure as follows (see full details in [12]). First, for each individual AU we performed a Pearson correlation between two binary vectors – the first vector recorded the presence vs. absence of the AU considered on each trial; the second vector recorded the responses of the participant on each corresponding trial (e.g., 'sadness'). For all AUs significantly correlated with a given emotion, we assigned a value of 1 (two-tailed $p < 0.05$) and 0 otherwise. This resulted in a $1\times$ 41-dimensional binary vector per emotion that details the AUs that are significantly correlated with the perception of that emotion for each individual participant. We did not analyze trials categorized as 'other' because they do not correspond to any specific emotion. Then, for each significant AU in the 1 $\times$ 41-dimensional binary vector, we computed an estimate of its temporal dynamics as follows. For each of the 7 temporal parameters, we performed an independent linear regression between the participant's intensity ratings (e.g., from 'very weak' to 'very strong') and all trials where they selected the emotion in question (e.g., 'sadness'). Thus, we computed a total of 12 dynamic facial expression models (2 participants $\times$ 6 emotions) each represented as a 1 $\times$ 41-dimensional binary vector detailing the significant AUs, plus 7 values detailing the temporal parameters of each significant AU. Computing these dynamic facial expressions in this way enables their reconstruction as stimuli for subsequent validation. To derive movies of the resulting dynamic facial expressions, we combined the significantly correlated AUs with their corresponding temporal parameters derived from the regression coefficients. Fig. 2*A* shows examples of the facial expressions that are currently installed in the social robot head (top row) and those derived using our reverse-engineering approach (bottom row) displayed as color-coded face maps – red indicates high AU amplitude, blue indicates low AU amplitude (see colorbar to right).

### D. Comparison of Recognition Accuracy with Existing Social Robot Facial Expressions

To compare the recognition accuracy of our reverse-engineered facial expressions and the social robot's current facial expressions, we asked a new group of participants

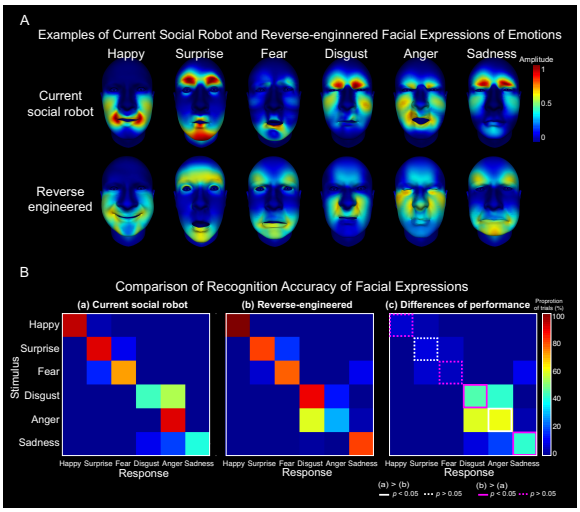


Figure 2. *A*. Examples of facial expressions currently installed in the social robot (top row) and those reverse-engineered using our methods (bottom row). Each facial expression is shown as a static face map. *B*. Recognition accuracy of the (a) current social robot facial expressions and (b) those derived using our methods, with their differences shown in (c).

to categorize each facial expression by emotion in a 6-alternative forced choice task. We recruited 10 white Westerners (5 female, mean age 24 years, SD = 3.18 years) using the same criteria as in *A. Modelling Dynamic Facial Expressions of Emotion*. Since high intensity facial expressions are perceived as less natural [21], we used the facial expressions derived from the medium intensity ratings. We displayed each facial expression model on each of the 7 face textures ('Default,' 'Male,' 'Female,' 'Obama,' 'iRobot,' 'Gabriel,' and 'Avatar') and repeated this procedure twice. We thus produced a total of 266 facial animations for each participant (2 reverse-engineered facial expressions $\times$ 6 emotions $\times$ 7 face textures $\times$ 2 repetitions + 7 current social robot facial expressions [2 happy facial expressions + 1 facial expression for each of other 5 emotions] $\times$ 7 face textures $\times$ 2 repetitions). On each experimental trial, the participant viewed a facial animation displayed on the social robot head and categorized it according to one of the six emotions. We played each facial animation only once for a duration of 2.25 seconds. Each participant categorized all 266 animations with trials presented in random order across the experiment. We used the same experimental conditions and equipment as in *A. Modelling Dynamic Facial Expressions of Emotion.*

To compute the recognition accuracy of the reverse-engineered and the current social robot facial expressions, we computed the proportion of correct responses for each emotion separately by pooling all trials across the face textures and participants. To identify any specific confusions between emotions, we also computed the proportion of incorrect responses distributed across the other emotion categories in the same way. Panels (a) and (b) in Fig.

2*B* show the results. Each cell of the color-coded matrices represents the proportion of correct trials (diagonal squares) or incorrect trials (off diagonal squares) for each emotion. Red indicates a high proportion of trials and blue indicates a low proportion (see colorbar to right). Panel (c) in Fig. 2*B* shows the absolute differences of performance between the reverse-engineered and the current social robot facial expressions. To determine any significant differences in accuracy between the reverse-engineered and current social robot facial expressions, we used a Monte Carlo simulation method to randomly shuffle the responses of each participant. We repeated this procedure 1000 times to derive a distribution of values for each emotion. We then determined statistical significance by comparing the human responses to 95% of the samples derived from shuffling the data (i.e., two-tailed test; $p < 0.05$). As shown by the squares outlined in pink in Fig 3(c), the reverse-engineered facial expressions elicited higher recognition accuracy for 'happy,' 'fear,' 'disgust' and 'sadness,' with 'disgust' and 'sadness' reaching statistical significance (solid line). The social robot's current facial expression of 'anger' was recognized with significantly higher accuracy (outlined in white) than the reverse-engineered facial expressions.

### *E. Comparison of Judgments of Huamanlike-ness with Existing Social Robot Facial Expressions*

Finally, we examined whether human users would judge our reverse-engineered facial expressions as more humanlike than the current social robot facial expressions. On each trial, we presented two facial expressions of the same emotion sequentially – one reverse-engineered, one current social robot – displayed on the same face texture, and asked participants to choose which one looked most humanlike. We included all facial expressions, textures, and pair combinations (including both temporal orders), blocked trials by emotion and face texture, and randomized the order of the blocks and the order of the trials within each block for each participant. Each participant completed a total of 196 trials (14 pairs of facial expressions of the same emotion [4 pairs for 'happy' + 2 pairs for each of the other emotions] $\times$ 2 temporal orders $\times$ 7 face textures). For each set of facial expressions and emotion separately, we computed the proportion of trials that humans perceived to be more humanlike by pooling all trials across face textures and participants. To determine whether the responses are significantly different from chance, we applied a Monte Carlo simulation-type method by randomly generating a set of responses (1000 iterations) and computing the proportion of trials assigned as more humanlike at each iteration. We determined statistical significance by comparing our experimental results to 95% of the randomly generated samples (i.e., $p < 0.05$). Fig. 3 shows the results. The white bars show that human users perceived the reverse-engineered facial expressions (white bars) as more humanlike than the social robot facial expressions (black bars) significantly more often than chance for 'happy,' 'surprise,' 'anger,' and 'sadness' (see red asterisks), with a slight advantage for 'fear' and 'disgust.'

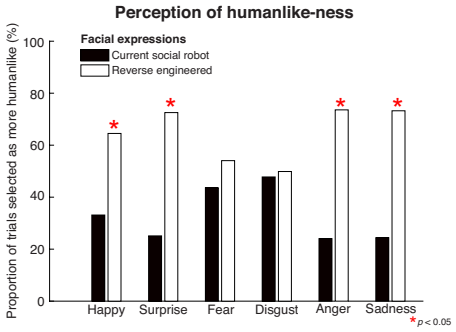


Figure 3. Proportion of trials perceived as more humanlike by human users. Black bars represent the current social robot facial expressions; white bars represent the reverse-engineered facial expressions. Red asterisks indicate statistical significance.

## IV. CONCLUSIONS

Here, we applied the psychophysical method of reverse correlation combined with subjective human perception to reverse engineer for the first time dynamic facial expression models of the six classic emotions directly on a social robot head. We show that human users categorize our reverse-engineered facial expressions of 'happy,' 'fear,' 'disgust,' and 'sadness' with higher accuracy, and are judged as humanlike more often, than the social robot's current facial expressions. These results therefore demonstrate the benefits of this data-driven approach to deriving psychologically valid dynamic facial expressions for social robots. One potential development that could further improve these facial expressions is the addition of dynamic textures, which produce the appearance of fine wrinkles on the face when certain AUs are activated – for example, Nose Wrinkler – and which are key to human emotion perception (e.g., [22], [23]). Our future work will therefore aim to transfer this animation display system to this social robot and other platforms including virtual humans (e.g., [24]) and robot heads with artificial skin (e.g., [25]).

In sum, our main aim is to demonstrate the generic power of a platform that combines data-driven psychophysical methods with subjective human perception to reverse-engineer psychologically valid dynamic facial expressions for social robotics. Though our demonstration is focused on the six classic emotions, we anticipate that our platform, with specific developments for compatibility with different social robot display systems, will equip social robots with a broader range of facial expressions such as those used during conversations (e.g., interested and confusion [26]), to convey pain or pleasure [27], or personality traits [28], thereby improving the quality of human-robot interactions.

## REFERENCES


[1] P. Ekman and W. V. Friesen, *Manual for the facial action coding system*: Consulting Psychologists Press, 1978.

[2] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, 2010, pp. 94-101.

[3] T. Hashimoto, S. Hitramatsu, T. Tsuji, and H. Kobayashi, "Development of the face robot SAYA for rich facial expressions," in *SICE-ICASE, International Joint Conference*, 2006, pp. 5423-5428.

[4] C. L. Breazeal, *Designing sociable robots*: MIT press, 2004.

[5] P. Ekman, E. R. Sorenson, and W. V. Friesen, "Pan-cultural elements in facial displays of emotion," *Science*, vol. 164, pp. 86-88, 1969.

[6] R. E. Jack, O. G. Garrod, and P. G. Schyns, "Dynamic Facial Expressions of Emotion Transmit an Evolving Hierarchy of Signals over Time," *Current Biology*, 2014.

[7] R. E. Jack, "Culture and facial expressions of emotion," *Visual Cognition*, vol. 21, pp. 1248-1286, 2013.

[8] N. L. Nelson and J. A. Russell, "Universality revisited," *Emotion Review*, vol. 5, pp. 8-15, 2013.

[9] H. A. Elfenbein and N. Ambady, "Is there an in-group advantage in emotion recognition?" 2002.

[10] I. Poggi and C. Pelachaud, "Performative facial expressions in animated faces," *Embodied conversational agents*, pp. 155-189, 2000.

[11] R. E. Jack, O. G. Garrod, H. Yu, R. Caldara, and P. G. Schyns, "Facial expressions of emotion are not culturally universal," *Proceedings of the National Academy of Sciences*, vol. 109, pp. 7241-7244, 2012.

[12] H. Yu, O. G. Garrod, and P. G. Schyns, "Perception-driven facial expression synthesis," *Computers and Graphics*, vol. 36, pp. 152-162, 2012.

[13] A. Ahumada and J. Lovell, "Stimulus features in signal detection.," *Journal of the Acoustical Society of America*, vol. 49, pp. 1751-1756, 1971.

[14] R. E. Jack, C. Crivelli, and T. Wheatley, "Data-Driven Methods to Diversify Knowledge of Human Psychology," *Trends in cognitive sciences*, vol. 22, pp. 1-5, 2018.

[15] R. E. Jack, R. Caldara, and P. G. Schyns, "Internal representations reveal cultural diversity in expectations of facial expressions of emotion," *Journal of Experimental Psychology: General*, vol. 141, p. 19, 2012.

[16] R. E. Jack and P. G. Schyns, "Toward a Social Psychophysics of Face Communication," *Annual Review of Psychology*, vol. 68, pp. 269-297, 2017.

[17] L. Ibrahimagic-Seper, A. Celebic, N. Petricevic, and E. Selimovic, "Anthropometric differences between males and females in face dimensions and dimensions of central maxillary incisors," *Medicinski glasnik*, vol. 3, pp. 58-62, 2006.

[18] E. Hall, "Distances in Man," *The Hidden Dimension*, pp. 101-129, 1966.

[19] J. De Leersnyder, B. Mesquita, and H. S. Kim, "Where do my emotions belong? A study of immigrants' emotional acculturation," *Personality and Social Psychology Bulletin*, vol. 37, pp. 451-463, 2011.

[20] S. Al Moubayed, J. Beskow, G. Skantze, and B. Granstrm, "Furhat: a back-projected human-like robot head for multiparty human-machine interaction," *Cognitive behavioural systems*, pp. 114-130, 2012.

[21] M. Ochs, R. Niewiadomski, and C. Pelachaud, "18 Facial Expressions of Emotions for Virtual Characters," *The Oxford Handbook of Affective Computing*, p. 261, 2015.

[22] F. W. Smith and P. G. Schyns, "Smile through your fear and sadness transmitting and identifying facial expression signals over a range of viewing distances," *Psychological Science*, vol. 20, pp. 1202-1208, 2009.

[23] M. L. Smith, G. W. Cottrell, F. Gosselin, and P. G. Schyns, "Transmitting and decoding facial expressions," *Psychological science*, vol. 16, pp. 184-189, 2005.

[24] J. Gratch, D. DeVault, G. M. Lucas, and S. Marsella, "Negotiation as a challenge problem for virtual humans," in *International Conference on Intelligent Virtual Agents*, 2015, pp. 201-215.

[25] T. Wu, N. J. Butko, P. Ruvulo, M. S. Bartlett, and J. R. Movellan, "Learning to make facial expressions," in *Development and Learning, 2009. ICDL 2009. IEEE 8th International Conference on*, 2009, pp. 1-6.

[26] C. Chen, O. Garrod, P. Schyns, and R. Jack, "The Face is the Mirror of the Cultural Mind," *Journal of vision*, vol. 15, pp. 928-928, 2015.

[27] C. Chen, C. Crivelli, O. Garrod, J.-M. Fernandez-Dols, P. Schyns, and R. Jack, "Facial Expressions of Pain and Pleasure are Highly Distinct," *Journal of Vision*, vol. 16, 2016.

[28] D. Gill, O. G. Garrod, R. E. Jack, and P. G. Schyns, "Facial movements strategically camouflage involuntary social signals of face morphology," *Psychological science*, vol. 25, pp. 1079-1086, 2014.